BERTOPIC: NEYRON TEMATIK MODELLASHTIRISH USULI

Narzillo Aloyev Raxmatilloyevich,
Tayanch doktorant
vip.alayev@gmail.com
ToshDO‘TAU

Annotatsiya. Tematik modellashtirish – bu katta matn to‘plamlarida yashirin tematik tuzilmalarni aniqlashga yordam beradigan tabiiy tilni qayta ishlash usulidir. Yashirin Dirixle taqsimoti (LDA) va manfiy bo‘lmagan matritsa faktorizatsiyasi (NMF) kabi an'anaviy yondashuvlar ko‘p yillar davomida tematik modellashtirish sohasida asosiy o‘rinlarni egallab kelishgan, ammo ular so‘zlar orasidagi semantik munosabatlarni e'tiborsiz qoldiradigan Bag-of-Words ga tayanganligi sababli turxli xil cheklovlarga duch kelishmoqda. BERTopic o‘z navbatida transformatorlarga asoslangan holda til modellarining semantikasini birlashtirish orqali klasterlash usullari va maxsus ishlab chiqilgan sinfga asoslangan TF-IDF protsedurasi bilan birlashtirilgan holda, tematik modellashtirishga yangi yondashuvni taqdim etadi. Ushbu kombinatsiya moslashuvchan bo‘lib adaptatsiyalashgan tizimni saqlagan holda, yanada interpretatsiya qilinadigan mavzularni yaratishga imkon beradi.

Annotation. Topic modeling is a natural language processing technique that helps to identify hidden thematic structures in large text collections all over the world and in any language. Traditional approaches such as LDA and NMF have been the mainstays of thematic modeling for many years, but they suffer from various limitations due to their reliance on “Bag-of-Words”, which ignores semantic relationships between words in texts. BERTopic, in turn, provides a new approach to thematic modeling by combining the semantics of language models based on transformers and a specially designed class-based TF-IDF procedure. This combination allows for the creation of more interpretable topics and understand of many texts in any language.

Аннотация. Тематическое моделирование – это метод обработки естественного языка, который помогает выявлять скрытые тематические структуры в больших текстовых коллекциях. Традиционные подходы, такие как скрытое распределение Дирихле (LDA) и неотрицательная матричная факторизация (NMF), на протяжении многих лет были основой тематического моделирования, но они сталкиваются с различными ограничениями из-за своей зависимости от «Bag-of-Words», который игнорирует семантические связи между словами. BERTopic, в свою очередь, представляет новый подход к тематическому моделированию, объединяя семантику языковых моделей на основе трансформаторов и специально разработанной процедурой TF-IDF на

основе классов. Такое сочетание позволяет создавать более интерпретируемые темы, сохраняя при этом гибкую и адаптируемую систему.

Kalit so‘zlar: BERTopic, BERT, TF-IDF, LDA, NMF, neyron tematik modellashtirish

Mavzuni modellashtirishning an’anaviy usullari tadqiqotchilarga katta hujjatlar to‘plamini tahlil qilishda yordam berishda asosiy rol o‘ynaydi, ammo ular sezilarli cheklovlarga ega:

1) **Bag-of-Words cheklovi**: LDA kabi modellar hujjatlarga kontekstual va semantik munosabatlarni e'tiborsiz qoldirib, tartibsiz so‘zlar to‘plami sifatida qaraydi.
2) **Semantik tushunchaning yetishmasligi:** An'anaviy modellar nozik ma'nolarni va so‘zlar orasidagi munosabatlarni tu’hunishda qiyinchiliklarga duch keladi.
3) **Chekli lug‘at:** Ko‘pgina an’anaviy usullar oldindan belgilangan lug‘atni talab qiladi, bu esa o‘z navbatida ularning moslashuvchanligini cheklab qo‘yadi.

Tabiiy tilni qayta ishlash sohasidagi so‘nggi yutuqlar, xususan, BERT kabi transformatorga asoslangan til modellari hajmga ega vektor bo‘shliqlaridan kontekstual ma'nolarni ajratib olish orqali matn tasvirini yangicha tarzda inqilob qildi. Tematik modellashtirishning bir nechta usullari ushbu qo‘shimchalardan foydalanishga harakat qildi, ammo ko‘pchiliklari mavzularning izchil tasvirlarini yaratishda hali ham qiyinchiliklarga duch kelmoqda. BERTopic hujjatni joylashtirish, klasterlash va mavzuni ko‘rsatish bosqichlarini ajratuvchi modulli tizimni amalga oshirish orqali ushbu muammolarni hal qiladi va uni turli til modellari va foydalanish holatlariga moslashtiradi.

BERTopic metodologiyasi uchta asosiy komponentlardan iborat:

1. **Hujjatlarning vektor tasviri**

Model hujjatlarni hajmga ega vektorli tasvirlarga aylantirish uchun Sentence-BERT (SBERT) dan foydalanadi. SBERT semantik o‘xshashlik vazifalari uchun nozik sozlangan oldindan o‘rgatilgan til modellaridan foydalanadi, bu o‘xshash ma’noga ega bo‘lgan hujjatlarni vektor fazosida bir-biriga yaqin joylashtirish imkonini beradi. Ushbu usulning moslashuvchanligi, MiniLM kabi kichikroq va MPNET kabi kattaroq va kuchliroq modellarga qadar oldindan o‘rgatilgan turli xil til modellaridan foydalanishga imkon beradi.

2. **Hujjatlarni klasterslash**

Hujjatlarning vektor tasviri yaratilganidan so‘ng BERT quyidagilarni amalga oshiradi:

a) PCA yoki t-SNE kabi alternativalarga qaraganda yuqori o‘lchamli ma'lumotlarning mahalliy va global xususiyatlarini yaxshi saqlaydigan UMAP (Uniform Manifold Approximation and Projection) yordamida o‘lchamlarni kamaytirish.

b) HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) yordamida klasterlash ya’ni, har xil zichlikdagi klasterlarni topadi. Ushbu yondashuv bir-biriga bog‘liq bo‘lmagan hujjatlarni har qanday klasterga biriktirilishining oldi olinadi hamda, mavzularning taqdim etilishi sifat jihatidan ancha yaxshilanadi.

UMAP uchun matematik formulani minimallashtirish sifatida quyidagicha ko‘rsatish mumkin:

$$\sum(i,j)\,[\,v_{ij} * log(w_{ij}) + (1 - v_{ij}) * log(1 - w_{ij})\,]$$

Bu yerda v_ij katta o‘lchamli o‘xshashlikni va w_ij kichik o‘lchamli o‘xshashlikni ifodalaydi.

3. Matnning ko‘rinishi

Aynan shu qismda BERTopic o‘zining eng innovatsion yo‘nalishini taqdim etadi: sinflarga asoslangan (class-based) TF-IDF yondashuvi.

An’anaviy TF-IDF hujjatdagi so‘zning hujjatlar to‘plamiga nisbatan ahamiyatini ifodalaydi. BERTopic ushbu kontseptsiyani hujjat darajasida emas, balki klaster darajasida ishlashga moslashtiradi. Sinflarga asoslangan TF-IDF quyidagi tartiblarni o‘z ichiga oladi:

a. **Hujjatlarni yig‘ish**: Klasterdagi barcha hujjatlar bitta “hujjat” ko‘rinishiga keltiriladi

b. **Sinflarga asoslangan atama chastotasi**: atamalarning har bir klasterda uchratilishining chastotasini hisoblash

c. **Hujjatning o‘zgartirilgan teskari chastotasi**: sinfga asoslangan holdagi ko‘rinishini sozlash

c-TF-IDF matematik formulasi quyida keltiriladi:

$$c\text{-}TF\text{-}IDF(t,c,C) = tf(t,c) * idf(t,C)$$

Bu yerda:

- tf(t,c) — c sinfdagi t terminning chastotasi
- idf(t,C) = log(1 + |C|/(1 + df(t,C)))
- |C| - sinflar soni
- df(t,C) - t atamasi mavjud bo‘lgan sinflar soni

Ushbu yondashuv bir qator afzalliklarni taqdim etadi:

- Mavzularni taqdim etishda barcha hujjatlar klasteri hisobga olinadi
- Muayyan mavzuga xos bo‘lgan so‘zlarga ustunlik beradi
- U ma’noli va so‘zlarga boy mavzular ko‘rinishini yaratadi.

c-TF-IDF tasvirlar generatsiya qilinganidan so‘ng, BERTopic o‘z navbatida har bir mavzu uchun eng yuqori ko‘rsatkichga ega n ta so‘zni tanlab oladi.

BERTopic o‘z doirasini dinamik mavzularni modellashtirish uchun kengaytiradi va mavzular vaqt o‘tishi bilan qanday rivojlanishini tahlil qilish imkonini beradi. Quyida yondashuvlar berilgan:

1) Mavzularning global tasvirini yaratish uchun BERTopic butun korpusga qo‘llaniladi
2) Global mavzu strukturasidan foydalangan holda har bir vaqt davri uchun mahalliy ko‘rinishni yaratadi
3) Mavzu evolyutsiyasini modellashtirish uchun har bir bosqichda yaratilgan c-TF-IDF matritsalaridan foydalanadi

Bu bizga har bir vaqt davri uchun butun bir modelni qayta sozlashni talab qilmasdan, vaqt o‘tishi bilan mavzuning tarqalishini va tarkib jihatidan o‘zgarishini kuzatish imkonini beradi.

BERTopic bir nechta ma’lumotlar to‘plamida, jumladan 20 NewsGroups, BBC News va Topicmodel.uz da baholanib ko‘rildi. Asosiy urg‘ular quyidagilarga mos keladi:

1) **Mavzuning ma’nolik darajasi**: BERTopic odatda turli ma’lumotlar to‘plamlari bo‘yicha yuqori ko‘rsatkichga ega, bu esa u yaratgan mavzular semantik jihatdan mazmunli ekanligini ko‘rsatadi.

2) **Mavzu xilma-xilligi**: BERTopic mavzular xilma-xilligida raqobatbardosh bo‘lsa-da, CTM (Kontekstlashtirilgan mavzu modeli) doimiy ravishda ustunlikka ega bo‘lib kelmoqda.

3) **Til modellari orasida barqarorlik belgisi**: BERTopicning ishlash mahsuldorligi SBERT til modellariga nisbatan barqarorligicha qolmoqda, bu uning mustahkamligi va moslashuvchanligini ko‘rsatadi.

4) **Dinamik mavzularni modellashtirishning samaradorligi**: BERTopic vaqt o‘tishi bilan mavzular evolyutsiyasini kuzatishda o‘zini yaxshi tomondan ko‘rsatadi.

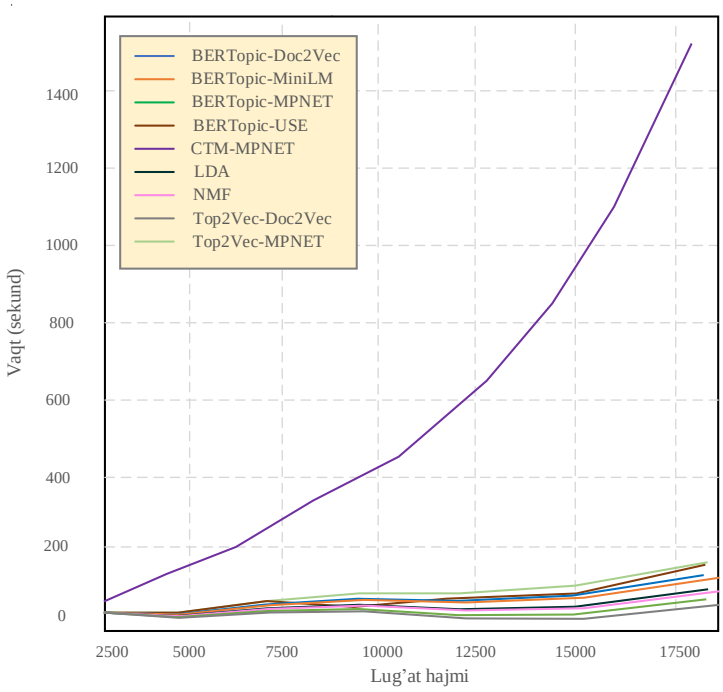
BERTopic tomonidan natija sifatida mavzular misoli quyidagicha ko‘rinishi mumkin:

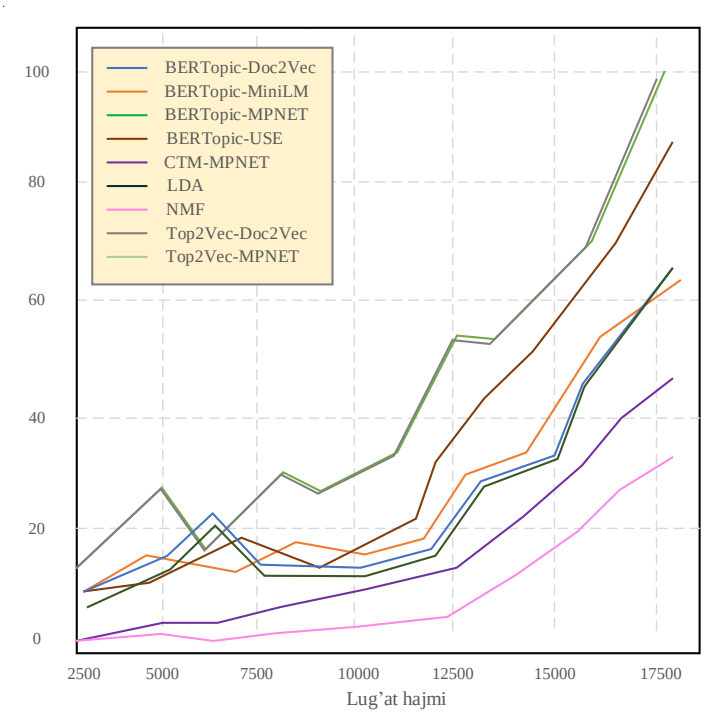
1-Mavzu: ['fazo‘, 'nasa', 'shatl', 'yer', 'start', 'missiya', 'orbita']

2-Mavzu: ['o‘yin', 'jamoa', 'o‘yinchi', 'sezon', 'xokkey', 'liga']

3-Mavzu: ['shifrlash', 'kalit', 'xavfsizlik', maxfiylik', 'davlat']

Ushbu mavzu ko‘rinishlari odatda an’anaviy mavzu modellari tomonidan ishlab chiqarilganidan ko‘ra ko‘proq interpretatsiyaga moyildir.





1-rasm: LDA va NMF kabi an'anaviy yondashuvlar bilan solishtirganda BERTopic ning hisoblash samaradorligini taqqoslash. Chapdagi grafik to‘liq bajarilish vaqtini taqqoslash, o‘ng grafik esa bajarilish vaqti shkalasining quyi qismiga qaratilgan.

1-rasmda ko‘rsatilganidek, BERTopic ishlash va hisoblash samaradorligi jihatidan yaxshi ko‘rsatkichlarni taklif qiladi. Til modelini tanlash o‘z navbatida jarayon vaqtiga sezilarli darajada ta’sir qiladi:

a) MiniLM kabi kichikroq modellardan foydalanish BERTopic-ni tezlik bo‘yicha odatiy usullar bilan raqobatbardosh qiladi.

b) MPNET kabi kattaroq modellar yaxshiroq semantik ma’no jihatidan yaxshiroq ko‘rsatkichlarni taqdim qiladi, lekin o‘ziga yarasha ko‘proq hisoblash resurslarini talab qiladi.

c) BERTopic odatda CTM dan tezroq, lekin odatiy LDA yoki NMF usullariga qaraganda sekinroq.

Hisoblash murakkabligi lug‘at hajmi bilan ortadi, ammo o‘sish tezligi ba'zi raqobatdosh yondashuvlarga nisbatan mutanosib ravishda boshqaruvga egadir.

BERTopicning sohaga qo‘shgan hissasi quyidagilardan iborat:

1) **Mavzular uyg‘unligi yaxshilandi**: Semantikaning vektor ko‘rinishi va sinfga asoslangan TF-IDF yondashuvidan foydalangan holda, BERTopic yanada izchil va izohlanadigan mavzularni ishlab chiqaradi.
2) **Moslashuvchanlik**: Modulli dizayn turli til modellari bilan oson integratsiya qilish va muayyan foydalanish holatlari uchun sozlash imkonini beradi.
3) **Natijaga erishish**: HDBSCAN-dan foydalanish BERTopic-ga hech qanday mavzuga tegishli bo‘lmagan hujjatlarni aniqlash imkonini beradi, bu esa mavzularning umumiy sifatini yaxshilaydi.
4) **Dinamik mavzu tahlili**: BERTopicning mavzu evolyutsiyasini kuzatish qobiliyati vaqtinchalik tahlil uchun yangi imkoniyatlarni ochadi.

Hayotiy qo‘llanmalar:

- Yangiliklar va ijtimoiy tarmoqlarda kontent tahlili
- Mijozlarning sharhlarini turkumlash
- Akademik tadqiqot yo‘nalishlarini tahlil qilish
- Ilmiy adabiyotlarni o‘rganish
- Siyosiy va huquqiy hujjatlarni tashkil etish

Xulosa o‘rnida aytish joizki, ushbu usul oqilona hisoblash samaradorligini saqlab qolgan holda an’anaviy tematik modellaridan ko‘ra ko‘proq mahsuldorlikka egadir. Til modellari rivojlanishda davom etar ekan, BERTopicning moslashuvchan tuzilmasi unga ushbu yutuqlarni o‘z ichiga olish imkonini beradi va bu uning tabiiy tillarni qayta ishlashning jadal rivojlanayotgan sohasida doimiy dolzarbligini ta'minlab turadi.

Foydalanilgan adabiyotlar:

1. Elov B., Alayev R., Aloyev N. (2024). Tematik modellashtirishning zamonaviy usullari. *Digital transformation and artificial intelligence*, *2*(1), 8–16. Retrieved from https://dtai.tsue.uz/index.php/dtai/article/view/v2i12
2. Elov.B., Alayev N. Matnlarini tematik modellashtirish va tasniflash usullari. barqarorlik va yetakchi tadqiqotlar onlayn ilmiy-amaliy jurnali. Vol. 3 No. 12 (2023). 263-276
3. Elov B., Aloyev N., Yuldashev A. (2023). SVD va NMF metodlari orqali tematik modellashtirish. O`zbekiston: til va madaniyat (Kompyuter lingvistikasi), 2023, 2(6). 55-66
4. Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, *6*(1). https://doi.org/10.14569/ijacsa.2015.060121
5. Tao, R., Wei, Y., & Yang, T. (2021). Metaphor Analysis Method Based on Latent Semantic Analysis. *Journal of Donghua University (English Edition)*, *38*(1). https://doi.org/10.19884/j.1672-5220.202010087
6. Darmalaksana, W., Slamet, C., Zulfikar, W. B., Fadillah, I. F., Maylawati, D. S. adillah, & Ali, H. (2020). Latent semantic analysis and cosine similarity for hadith search engine. *Telkomnika (Telecommunication Computing Electronics and Control)*, *18*(1). https://doi.org/10.12928/TELKOMNIKA.V18I1.14874
7. Ke, Z. T., & Wang, M. (2022). Using SVD for Topic Modeling. *Journal of the American Statistical Association*.
8. https://doi.org/10.1080/01621459.2022.2123813
9. Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, *54*(10). https://doi.org/10.1145/3507900
10. Kherwa, P., & Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(24). https://doi.org/10.4108/eai.13-7-2018.159623